arXiv:2504.01990v1 [cs.AI] 31 Mar 2025

Advances and Challenges in Foundation Agents

From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems

Bang Liu[2,3,20*†], **Xinfeng Li**[4*], **Jiayi Zhang**[1,10*], **Jinlin Wang**[1*], **Tanjin He**[5*], **Sirui Hong**[1*], **Hongzhang Liu**[6*], **Shaokun Zhang**[7*], **Kaitao Song**[8*], **Kunlun Zhu**[9*], **Yuheng Cheng**[1*], **Suyuchen Wang**[2,3*], **Xiaoqiang Wang**[2,3*], **Yuyu Luo**[10*], **Haibo Jin**[9*],**Peiyan Zhang**[10], **Ollie Liu**[11], **Jiaqi Chen**[1], **Huan Zhang**[2,3], **Zhaoyang Yu**[1], **Haochen Shi**[2,3], **Boyan Li**[10], **Dekun Wu**[2,3], **Fengwei Teng**[1], **Xiaojun Jia**[4], **Jiawei Xu**[1], **Jinyu Xiang**[1], **Yizhang Lin**[1], **Tianming Liu**[14], **Tongliang Liu**[6], **Yu Su**[15], **Huan Sun**[15], **Glen Berseth**[2,3,20], **Jianyun Nie**[2], **Ian Foster**[5], **Logan Ward**[5], **Qingyun Wu**[7], **Yu Gu**[15], **Mingchen Zhuge**[16], **Xiangru Tang**[12], **Haohan Wang**[9], **Jiaxuan You**[9], **Chi Wang**[19], **Jian Pei**[17†], **Qiang Yang**[10,18†], **Xiaoliang Qi**[13†], **Chenglin Wu**[1*†]

[1]MetaGPT, [2]Université de Montréal, [3]Mila - Quebec AI Institute, [4]Nanyang Technological University, [5]Argonne National Laboratory, [6]University of Sydney, [7]Penn State University, [8]Microsoft Research Asia, [9]University of Illinois at Urbana-Champaign, [10]The Hong Kong University of Science and Technology, [11]University of Southern California, [12]Yale University, [13]Stanford University, [14]University of Georgia, [15]The Ohio State University, [16]King Abdullah University of Science and Technology, [17]Duke University, [18]The Hong Kong Polytechnic University, [19]Google DeepMind, [20]Canada CIFAR AI Chair

Abstract

The advent of large language models (LLMs) has catalyzed a transformative shift in artificial intelligence, paving the way for advanced intelligent agents capable of sophisticated reasoning, robust perception, and versatile action across diverse domains. As these agents increasingly drive AI research and practical applications, their design, evaluation, and continuous improvement present intricate, multifaceted challenges. This survey provides a comprehensive overview, framing intelligent agents within a modular, brain-inspired architecture that integrates principles from cognitive science, neuroscience, and computational research. We structure our exploration into four interconnected parts. First, we delve into the **modular foundation of intelligent agents**, systematically mapping their cognitive, perceptual, and operational modules onto analogous human brain functionalities, and elucidating core components such as memory, world modeling, reward processing, and emotion-like systems. Second, we discuss **self-enhancement and adaptive evolution mechanisms**, exploring how agents autonomously refine their capabilities, adapt to dynamic environments, and achieve continual learning through automated optimization paradigms, including emerging AutoML and LLM-driven optimization strategies. Third, we examine **collaborative and evolutionary multi-agent systems**, investigating the collective intelligence emerging from agent interactions, cooperation, and societal structures, highlighting parallels to human social dynamics. Finally, we address the critical imperative of **building safe, secure, and beneficial AI systems**, emphasizing intrinsic and extrinsic security threats, ethical alignment, robustness, and practical mitigation strategies necessary for trustworthy real-world deployment. By synthesizing modular AI architectures with insights from different disciplines, this survey identifies key research gaps, challenges, and opportunities, encouraging innovations that harmonize technological advancement with meaningful societal benefit. The project's Github link is: https://github.com/FoundationAgents/awesome-foundation-agents.

[*] Major Contribution. Work in progress.

[†] Corresponding authors: Bang Liu (bang.liu@umontreal.ca), Jian Pei (j.pei@duke.edu), Qiang Yang (qyang@cse.ust.hk), Xiaoliang Qi (xlqi@stanford.edu), Chenglin Wu (alexanderwu@deepwisdom.ai)

Preface

Large language models (LLMs) have revolutionized artificial intelligence (AI) by demonstrating unprecedented capabilities in natural language and multimodal understanding, as well as reasoning and generation. These models are trained on vast datasets, and they exhibit emergent abilities such as reasoning, in-context learning, and even rudimentary planning. While these models represent a major step forward in realizing intelligent machines, they themselves do not yet fully embody all the capabilities of an intelligent being. Since the early days of artificial intelligence, AI researchers have long been on a quest for a truly "intelligent" system that can learn, plan, reason, sense, communicate, act, remember, and demonstrate various human-like abilities and agility. These beings, known as intelligent agents, should be able to think both long-term and short-term, perform complex actions, and interact with humans and other agents. LLMs are an important step towards realizing intelligent agents, but we are not there yet.

This manuscript provides a comprehensive overview of the current state of the art of LLM-based intelligent agents. In the past, there have been numerous research papers and books on intelligent agents, as well as a flurry of books on LLMs. However, there has scarcely been comprehensive coverage of both. While LLMs can achieve significant capabilities required by agents, they only provide the foundations upon which further functionalities must be built. For example, while LLMs can help generate plans such as travel plans, they cannot yet generate fully complex plans for complex and professional tasks, nor can they maintain long-term memories without hallucination. Furthermore, their ability to perform real-world actions autonomously remains limited. We can view LLMs as engines, with agents being the cars, boats, and airplanes built using these engines. In this view, we naturally seek to move forward in designing and constructing fully functioning intelligent agents by making full use of the capabilities provided by LLMs.

In this engine-vehicle analogy of the interplay between LLMs and agents, we naturally ask: How much of the capabilities of intelligent agents can current LLM technologies provide? What are the functions that cannot yet be realized based on current LLM technologies? Beyond LLMs, what more needs to be done to have a fully intelligent agent capable of autonomous action and interaction in the physical world? What are the challenges for fully integrated LLM-based agents? What additional developments are required for capable, communicative agents that effectively collaborate with humans? What are the areas that represent low-hanging fruits for LLM-based agents? What implications will there be for society once we have fully intelligent LLM-based agents, and how should we prepare for this future?

These questions transcend not only the engineering practice of extending current LLMs and agents but also raise potential future research directions. We have assembled frontier researchers from AI, spanning from LLM development to agent design, to comprehensively address these questions. The book consists of four parts. The first part presents an exposition of the requirements for individual agents, comparing their capabilities with those of humans, including perception and action abilities. The second part explores agents' evolution capabilities and their implications on intelligent tools such as workflow management systems. The third part discusses societies of agents, emphasizing their collaborative and collective action capabilities, and the fourth part addresses ethical and societal aspects, including agent safety and responsibilities.

This book is intended for researchers, students, policymakers, and practitioners alike. The audience includes non-AI readers curious about AI, LLMs, and agents, as well as individuals interested in future societies where humans co-exist with AI. Readers may range from undergraduate and graduate students to researchers and industry practitioners. The book aims not only to provide answers to readers' questions about AI and agents but also to inspire them to ask new questions. Ultimately, we hope to motivate more people to join our endeavor in exploring this fertile research ground.

Contents